08/366083

MIT6834.APL
LMW4
LMW/ach
December 28, 1994

PATENT APPLICATION
Docket No.: MIT6834

CHIMERIC TRANSCRIPTION FACTORS

Government Support

Work described herein was supported by grants PO1-CA42063, CDR-8803014 and P30-CA14051 from the U.S. Public Health Service/National Institutes of Health, National Science Foundation and National Cancer Institute, respectively. The U.S. Government has certain rights in the invention. Work described herein was also supported by the Howard Hughes Medical Institute.

Background of the Invention

DNA-binding proteins, such as transcription factors, are critical regulators of gene expression. For example, transcriptional regulatory proteins are known to play a key role in systems evolved by cells to convert extracellular signals into altered gene expression (Curran and Franza, Cell 55:395-397 (1988)). DNA-binding proteins also play critical roles in the control of cell growth and in the expression of viral and bacterial genes.

There have been attempts made to change the specificity of DNA-binding proteins. The existing strategies primarily rely on making mutations in these proteins at sites important for DNA-recognition (Rebar and Pabo, Science 263:671-673 (1994), Jamieson et al., Biochemistry 33:5689-5695 (1994), Suckow et al., Nucleic Acids Research 22(12):2198-2208 (1994)). This strategy may not be efficient or possible with many DNA-binding domains because of limitations imposed by their three-dimensional structure and mode of docking to DNA. Therefore, it is desirable to have a strategy which can utilize many different DNA-binding domains and can combine them as required for DNA recognition and gene regulation.

Summary of the Invention

This invention pertains to proteins with novel nucleic acid binding specificities, particularly to proteins which bind DNA and are comprised of two or more separate DNA-binding domains which do not occur together in the same arrangement in nature (i.e., are comprised of two or more separate DNA-binding domains which a) do not occur together in nature; b) do not occur together in nature in the order in which they are present in a chimeric protein of the present invention; or c) do not occur together in nature with the spacing that is present in a chimeric protein of the present invention). In addition, the proteins of the present invention display DNA-binding specificity that is clearly distinct from that of the component DNA-binding domains; that is, they prefer binding the entire composite nucleotide sequence to binding a portion thereof. Two criteria suggest which arrangements of domains are suitable for combination in a chimeric protein which binds DNA: (1) lack of collision between component domains, and (2) consistent positioning of the carboxyl- and amino-terminal regions of each domain. Because of their ability to bind DNA normally bound by two or more separate DNA-binding domains which do not occur together in the same arrangement in nature, the proteins of the present invention are referred to as chimeric DNA-binding proteins.

The binding specificity of the chimeric DNA-binding proteins makes them particularly useful because they have DNA-binding properties distinct from those of known proteins. The chimeric proteins prefer to bind the composite nucleic acid sequence and, thus, alter expression of genes controlled by the regulatory element which contains the selected composite nucleic acid sequence. Preferably, the chimeric proteins do not bind to a significant extent the DNA bound by the component domains

of the chimera, and, thus, do not alter normal cellular gene expression.

In one embodiment of the present invention, the chimeric proteins bind selected nucleic acids within DNA or RNA and, as a result, mark or flag the selected DNA or RNA sequence, which can be identified and/or isolated from the DNA using known methods. In this respect, they act in a manner similar to restriction enzymes, in that they recognize DNA or RNA at a selected nucleic acid sequence, thus marking that sequence where ever it occurs in DNA or RNA with which the chimeric proteins are contacted. Unlike restriction enzymes, chimeric DNA-binding or RNA-binding proteins do not cut or fragment the DNA or RNA at the nucleic acids they recognize. Chimeric proteins used for this purpose can be labelled, e.g., radioactively or with an affinity ligand or epitope tag such as the GST peptide, and, thus, the location of DNA or RNA to which they bind can be identified easily. Because of the binding specificity of the chimeric proteins, DNA or RNA to which binding occurs must include either the nucleotide sequence which the chimeric proteins have been designed to recognize or the nucleotide sequences recognized by the component DNA-binding domains. Optimally, the chimeric protein will not efficiently recognize the nucleotide sequence recognized by the component DNA-binding domains. Standard methods, such as DNA cloning and dideoxysequencing, can be used to determine the nucleotide sequence to which the chimeric protein is bound.

A variety of components can be included in a chimeric protein of the present invention. For example, two known DNA-binding domains can be combined to produce a chimeric protein which is a novel DNA-binding protein. The chimeric binding protein displays DNA-binding specificity that is clearly distinct from that of the component DNA-binding domains of the chimeric protein; it greatly prefers to bind

the composite nucleic acid sequence. The component DNA-binding domains of the chimeric proteins are joined, either directly or through one amino acid or through a short polypeptide (two or more amino acids) to form a continuous polypeptide. Additional domains with desired properties can optionally be included in the chimeric proteins of the present invention.

In another embodiment of the present invention, proteins with novel DNA-binding specificities are transcription factors which bind certain DNA sequences and regulate gene expression. The transcription factors do not regulate expression of genes controlled by regulatory elements containing the individual DNA sequences bound by the component DNA-binding domains, but instead regulate expression of genes whose regulatory elements contain a composite DNA sequence which comprises all or a part of the nucleic acid sequences bound by the chimeric transcription factor components.

In a particular embodiment, the chimeric protein is a DNA-binding protein comprising at least one homeodomain, a short polypeptide linker and at least one zinc finger. The chimeric protein of this embodiment comprises, for example, zinc finger 1 or zinc finger 2 of Zif268, an amino acid or a short polypeptide, and the Oct-1 homeodomain. Alternatively, the chimeric protein comprises zinc fingers 1 and 2 of Zif268, a short linker, such as a glycine-glycine-arginine-arginine polypeptide, and the Oct-1 homeodomain. The latter chimeric protein, designated ZFHD1, is described below.

In an alternate embodiment, the novel DNA-binding protein is a chimeric zinc finger-basic-helix-loop-helix protein. For example, such a chimeric protein comprises fingers 1 and 2 of Zif268 and the MyoD bHLH region, joined by a linker which spans approximately 9.5 Å between the

carboxyl-terminal region of finger 2 and the amino-terminal region of the basic region of the bHLH domain.

In another embodiment, the novel chimeric protein is a zinc finger-steroid receptor chimera. Such a chimeric protein comprises, for example, fingers 1 and 2 of Zif268 and the DNA-binding domains of the glucocorticoid receptor, joined at the carboxyl-terminal region of finger 2 and the amino-terminal region of the DNA-binding domain of the glucocorticoid receptor by a linker which spans approximately 7.4 Å.

This invention also pertains to chimeric transcription factors. In one embodiment, the chimeric protein is a novel transcription factor comprising at least two DNA-binding domains, each of which is a DNA-binding polypeptide, such as a homeodomain and a zinc finger, joined by an amino acid or a short polypeptide linker, and an activation domain. For example, the chimeric protein ZFHD1-VP16, described below, comprises zinc fingers 1 and 2 of Zif268, a glycine-glycine-arginine-arginine polypeptide linker, the Oct-1 homeodomain, and the Herpes Simplex Virus VP16 activation domain.

The present invention further relates to a method of identifying DNA binding sites which are specifically bound by a chimeric protein of the present invention. For example, a chimeric protein which comprises a first DNA-binding domain which recognizes one nucleic acid sequence (e.g., a zinc finger) and a second DNA-binding domain which recognizes another nucleic acid sequence (e.g., a homeodomain), which second domain is not normally (in nature) present in the same DNA binding protein with the first DNA-binding domain, can be used to identify cellular DNA which includes the "composite" binding site (i.e., a site which is bound by the chimeric protein). Alternatively, a chimeric protein which comprises two or more DNA-binding domains which do not occur together in

nature in the order in which they are present in the chimeric protein or two or more DNA-binding domains which do not occur in nature with the same spacing with which they occur in the chimeric protein can be used.

The present invention also relates to a method of targeted regulation of specific genes. In one embodiment of the present method, the chimeric protein is a transcription factor which is a transcription activator and which comprises at least two DNA-binding domains and one or more activation domains. The chimeric transcription factor specifically binds to DNA target sites and, as a result, positively regulates transcription of a gene which is under the control of a regulatory element whose nucleic acid sequence includes a nucleic acid sequence bound by the chimeric protein transcription factor. The nucleic acid sequence recognized by the chimeric transcription factor can be present in the regulatory element as it occurs in nature (in the cells as obtained) or can be introduced into cells using known methods.

The transcription factor specifically binds to its composite sequence, which is contained in a regulatory element, such as a promoter, which controls a gene, and thereby positively regulates (e.g., turns on or increases) the rate of transcription of the gene. A non-native nucleic acid sequence comprising the nucleic acid sequence recognized by the chimeric protein may be inserted into a regulatory element of a gene in order to correct a naturally-occurring error in transcription or to improve or control the transcription of the gene through binding of a chimeric protein designed to recognize the inserted sequence. Alternatively, an additional copy of the gene may be introduced along with a novel regulatory region which contains one or more binding sites for the chimeric protein. In this way, chimeric proteins of the present

invention may be transcription factors whose DNA-binding regulates transcription of a gene.

In another embodiment, the chimeric protein is a transcription repressor which comprises at least two DNA-binding domains. The chimeric transcription repressor specifically binds to DNA target sites and thereby negatively regulates transcription of a gene which is under the control of a regulatory element whose nucleic acid sequence includes the nucleic acid sequence bound by the chimeric protein transcription repressor. The binding of the chimeric transcription repressor inhibits binding of a naturally-occurring transcription factor and thereby inhibits transcription of the gene.

Alternatively, the repressor may further comprise a repression domain. The binding of the repressor to its composite DNA sequence negatively regulates (e.g., turns off or decreases) transcription of the gene.

In an alternate embodiment, known mutational strategies can be used to expand the range of possible DNA-binding specificities of the chimeric proteins. For instance, the amino acid sequence of the DNA-binding domains of the chimeric protein may be altered through mutational procedures to allow the chimeric protein to bind different DNA sites from those bound by the unaltered chimera. Mutations or variations in the linker region may also be used to optimize the DNA-binding specificity or utility of the chimeric protein in a particular application or process.

This invention has application to several areas, including virtually any utility for which recognition of specific nucleic acid sequences is critical. For instance, the present invention is useful for gene regulation; that is, the novel DNA-binding chimeric proteins can be used for activation or repression of specific genes to control gene expression for production of recombinant gene products,

such as in cell culture. The present invention is also useful for gene therapy purposes, to correct or compensate for abnormal gene expression and control the expression of disease-causing gene products. The invention may also be used in gene therapy to increase the expression of a deficient gene product or decrease expression of a product which is overproduced or overactive.

The present invention also has utility for the manipulation of gene expression, e.g., the control of gene expression in a transgenic organism for protein production. The chimeric proteins of the present invention can also be used to identify specific rare DNA sequences for use as markers in gene mapping. Domains can be attached to the chimeric proteins which would cleave adjacent DNA, thus potentially generating a new series of endonuclease proteins.

Chimeric DNA-binding proteins of the present invention can also be used to induce or stabilize loop formation in DNA or to bring together or hold together DNA sites on two or more different molecules.

Brief Description of the Drawings

Figure 1A-C illustrates selection by ZFHD1 of a hybrid binding site from a pool of random oligonucleotides. Figure 1A is a graphic representation of the structure of the ZFHD1 chimeric protein used to select binding sites. (SEQ ID NO:32) The underlined residues are from the Zif268-DNA and Oct-1-DNA crystal structures and correspond to the termini used in computer modeling studies. The linker contains two glycines, which were included for flexibility and to help span the required distance between the termini of the domains, and the two arginines that are present at positions -1 and 1 of the Oct-1 homeodomain. A glutathione S-transferase domain (GST) is joined to the amino-terminus of zinc finger 1. Figure 1B shows the nucleic acid

sequences (SEQ ID NOS. 1-16, respectively) of 16 sites isolated after four rounds of binding site selection. These sequences were used to determine the consensus binding sequence (5'-TAATTANGGGNG-3', SEQ ID NO: 17) of ZFHD1. Figure 1C shows the alternative possibilities for homeodomain binding configurations suggested by the consensus sequence; Mode 1 was determined to be the correct optimal configuration for ZFHD1. The letter "N" at a position indicates that any nucleotide can occupy that position.

Figure 2A-C is an autoradiograph illustrating the DNA-binding specificity of ZFHD1, the Oct-1 POU domain and the three zinc fingers from Zif268. The probes used are listed at the top of each set of lanes, and the position of the protein-DNA complex is indicated by the arrow.

Figure 3 is a graphic representation of the regulation of promoter activity in vivo by ZFHD1. The expression vector encoded the ZFHD1 protein fused to the carboxyl-terminal 81 amino acids of VP16 (+ bars), and the empty expression vector Rc/CMV was used as control (- bars). Bar graphs represent the average of three independent trials. Actual values and standard deviation reading from left to right are: 1.00 ± .05, 3.30 ± .63; 0.96 ± .08, 42.2 ± 5.1; 0.76 ± .07, 2.36 ± .34; 1.22 ± .10, 4.22 ± 1.41. Fold induction refers to the level of normalized activity obtained with the ZFHD1-VP16 expression construct divided by that obtained with Rc/CMV.

Detailed Description of the Invention

This invention pertains to chimeric proteins which bind a composite DNA sequence; the chimeric protein comprises at least two domains, each of which is a DNA-binding polypeptide which binds a sequence which is a part of the composite DNA sequence.

In a particular embodiment, the chimeric proteins are transcription factors which may contain at least one

regulatory domain in addition to the DNA-binding domains. As used herein, the term "transcription factor" is defined as any protein that regulates transcription, and includes regulators that have a positive or a negative effect on transcription initiation or progression. Transcription factors may optionally contain one or more regulatory domains. As used herein, the term "regulatory domain" is defined as any domain which regulates transcription, and includes both activation domains and repression domains. As used herein, the term "activation domain" means a domain in a transcription factor which positively regulates (turns on or increases) the rate of gene transcription. The term "repression domain" means a domain in a transcription factor which negatively regulates (turns off or decreases) the rate of gene transcription. The nucleic acid sequence bound by a transcription factor is typically DNA required for expression or activity of a gene, such as within a promoter or regulatory element. However, sufficiently tight binding to nucleotides at other locations, e.g., within the coding sequence, can also be used to regulate gene expression.

The chimeric proteins of the present invention are comprised of an amino acid sequence which binds a first DNA sequence (e.g., sequence A) and an amino acid sequence which binds a second DNA sequence (e.g., sequence B). The amino acid sequences are DNA-binding domains which do not normally occur together in the same arrangement as they occur in a chimeric protein of the present invention. That is, the DNA-binding domains of the chimeric proteins do not occur together in the same protein in nature (i.e., are domains which occur in nature in separate proteins); do not occur together in nature in the order or orientation in which they occur in the chimeric protein; or do not occur together in nature with the same spacing as the spacing present in the chimeric protein. As used herein, a domain

of the chimeric proteins of the present invention may correspond to a subdomain of a natural DNA-binding domain, or may itself contain more than one natural DNA-binding domain.

The two amino acid sequences are linked (joined), either directly or through an amino acid linker, to form a chimeric protein which binds a third DNA sequence (e.g., sequence C, which typically contains sequence A and sequence B as subsequences). The chimeric protein displays DNA-binding specificity that is clearly distinct from that of the component DNA-binding domains, and greatly prefers to bind the composite sequence (e.g., sequence C) as opposed to the individual sequences (e.g., sequence A or sequence B). Amino acid sequences representing other domains, e.g., activation domains, repression domains, endonuclease cleavage domains or domains allowing interaction with other cellular components, may optionally be included in the chimeric protein.

The amino acids which comprise the chimeric proteins of the present invention may be naturally-occurring amino acids or modified amino acids (which do not occur in nature). The chimeric protein may include more than two DNA-binding domains. It may also include more than one linker, or include no linker, as appropriate to join the selected domains. The composite nucleic acid sequence recognized by the chimeric protein may include all or a portion of the sequences bound by the component amino acids. However, the chimeric protein displays a binding specificity that is distinct from the binding specificity of the amino acid components of the chimeric protein.

A structure-based strategy of fusing known DNA-binding modules has been used to design transcription factors with novel DNA-binding specificities. In order to visualize how certain DNA-binding domains might be fused to other DNA-

binding domains, computer modeling studies have been used to superimpose and align various protein-DNA complexes.

Two criteria suggest which alignments of DNA-binding domains have potential for combination in a chimeric protein which binds DNA: (1) lack of collision between domains, and (2) consistent positioning of the carboxyl- and amino-terminal regions of the domains, i.e., the domains must be oriented such that the carboxyl-terminal region of one polypeptide can be joined to the amino-terminal region of the next polypeptide, either directly or by a linker (indirectly). Domains positioned such that only the two amino-terminal regions are adjacent to each other or only the two carboxyl-terminal regions are adjacent to each other are not suitable for inclusion in the chimeric proteins of the present invention. When detailed structural information about the protein-DNA complexes is not available, it may be necessary to experiment with various endpoints, and more biochemical work may be necessary to characterize the DNA-binding properties of the chimeric proteins. This optimization can be performed using known techniques. Virtually any domains satisfying the above-described criteria are candidates for inclusion in the chimeric protein. Alternatively, non-computer modeling may also be used.

The amino acid sequences of the DNA-binding domains can be selected from any suitable DNA-binding proteins; a variety of DNA-binding proteins and their amino acid sequences are known. For example, DNA-binding domains of DNA-binding proteins with the helix-turn-helix structural design, including, but not limited to, MAT α1, MAT α2, MAT a1, Antennapedia, Ultrabithorax, Engrailed, Paired, Fushi tarazu, HOX, Unc86, Oct1, Oct2 and Pit, can be selected as components of a novel chimeric transcription factor. Other DNA-binding domains which can also be used are those from the zinc finger proteins, such as Zif268, SWI5, Krüppel and

Hunchback, as well as the steroid receptors. DNA-binding proteins with the helix-loop-helix structural design, such as Daughterless, Achaete-scute (T3), MyoD, E12 and E47, are also sources of useful DNA-binding domains, as are other helical motifs like the leucine-zipper, which includes GCN4, C/EBP, c-Fos/c-Jun and JunB. The amino acid sequences of the DNA-binding domains may be naturally-occurring or non-naturally-occurring (or modified).

The choice of DNA-binding domains utilized may be influenced by a number of considerations, including what species, system and cell type is targeted, which domains are shown by modeling to be feasible for incorporation in a chimeric protein and the desired application or utility. The choice of DNA-binding domains utilized will also be influenced by the individual DNA sequence specificity of the domain and the ability of the domain to interact with other proteins or to be influenced by a particular cellular regulatory pathway. Preferably, the distance between domain termini is relatively short to facilitate use of the shortest possible linker or no linker. The DNA-binding domains can be isolated from a naturally-occurring protein, or may be a synthetic molecule based in whole or in part on a naturally-occurring domain.

The homeodomain is a highly conserved DNA-binding domain which has been found in hundreds of transcription factors (Scott et al., Biochim. Biophys. Acta 989:25-48 (1989) and Rosenfeld, Genes Dev. 5:897-907 (1991)). The regulatory function of a homeodomain protein derives from the specificity of its interactions with DNA and presumably with components of the basic transcriptional machinery, such as RNA polymerase or accessory transcription factors (Laughon, Biochemistry 30(48):11357 (1991)).

The zinc finger motif, of the type first discovered in transcription factor IIIA (Miller et al., EMBO J. 4:1609 (1985)), offers an attractive framework for studies of

transcription factors with novel DNA-binding specificities. The zinc finger is one of the most common eukaryotic DNA-binding motifs (Jacobs, EMBO J. 11:4507 (1992)), and this family of proteins can recognize a diverse set of DNA sequences (Pavletich and Pabo, Science 261:1701 (1993)). Crystallographic studies of the Zif268-DNA complex and other zinc finger-DNA complexes show that residues at four positions within each finger make most of the base contacts, and there has been some discussion about rules that may explain zinc finger-DNA recognition (Desjarlais and Berg, PNAS 89:7345 (1992) and Klevit, Science 253:1367 (1991)). However, studies have also shown that zinc fingers can dock against DNA in a variety of ways (Pavletich and Pabo (1993) and Fairall et al., Nature 366:483 (1993)).

In one embodiment of the present invention, the chimeric protein includes at least two distinct DNA-binding domains (domains which are heterologous to one another). These two domains can be, for instance, the DNA-binding domains selected from the homeodomain family and the zinc finger family. As used herein, the phrase "zinc finger" refers to zinc finger domains that are homologous to those domains originally discovered in TFIIIA. For example, as described herein, a chimeric protein, in which the DNA-binding domains are the Oct-1 homeodomain and zinc fingers 1 and 2 of Zif268, has been produced and shown to bind a composite DNA sequence (SEQ ID NO: 17) which includes the nucleic acid sequences bound by the relevant portion of the two component DNA-binding proteins.

In another embodiment of the invention, the chimeric protein is a transcription factor which comprises at least two DNA-binding domains and at least one activation domain. For instance, as described herein, a chimeric protein (ZFHD1-VP16), in which the DNA-binding domains are the Oct-1 homeodomain and zinc fingers 1 and 2 of Zif268, and the

activation domain is the Herpes Simplex Virus VP16 activation domain, has been produced and shown to regulate transcription of a gene (the luciferase gene) in vivo.

The selected domains are joined into a continuous chimeric protein either directly or using one amino acid or a short polypeptide linker (indirectly). The linker may be any amino acid sequence that results in linkage of the component domains such that they retain the ability to bind their respective nucleotide sequences. Preferably the design should involve an arrangement of domains which requires the linker to span a relatively short distance. A distance of less than about 10 Å, but depending upon the selected DNA-binding domains and the configuration, the linker may span a distance of up to about 50 Å. For instance, the ZFHD1 protein contains a glycine-glycine-arginine-arginine linker which joins the carboxyl-terminal region of zinc finger 2 to the amino-terminal region of the Oct-1 homeodomain.

Within the linker, the amino acid sequence may be varied based on the preferred characteristics of the linker as revealed by modeling. For instance, in addition to a desired length, modeling studies may show that side groups of certain nucleotides or amino acids may interfere with binding of the protein. The primary criterion is that the linker join the DNA-binding domains in such a manner that they retain their ability to bind their respective DNA sequences, and thus a linker which interferes with this ability is undesirable. A desirable linker should also be able to constrain the relative three-dimensional positioning of the domains so that only certain composite sites are recognized by the chimeric protein. Other considerations in choosing the linker include flexibility of the linker, charge of the linker and selected binding domains, and presence of some amino acids of the linker in the naturally-occurring domains. The linker can also be

designed such that residues in the linker contact DNA, thereby influencing binding affinity or specificity, or to interact with other proteins. For example, a linker may contain an amino acid sequence which can be recognized by a protease so that the activity of the chimeric protein could be regulated by cleavage. In some cases, particularly when it is necessary to span a longer distance between the two DNA-binding domains or when the domains must be held in a particular configuration, the linker may optionally contain an additional folded domain.

Transcription factors can be tested for activity in vivo using a simple assay (F.M. Ausubel et al., Eds., Current Protocols in Molecular Biology (John Wiley & Sons, New York, 1994); de Wet et al., Mol. Cell Biol. 7:725 (1987)). The in vivo assay requires a plasmid containing the gene encoding the transcription factor in a form in which the protein will be expressed at suitable levels. The assay also requires a plasmid containing a test gene, e.g., the luciferase gene, the chloramphenicol acetyl transferase (CAT) gene or the human growth hormone (hGH) gene, with a binding site for the transcription factor. The two plasmids are introduced together into host cells whose genome lacks both the gene encoding the transcription factor and the test gene. A second group of cells, which also lack both the gene encoding the transcription factor and the test gene, serves as the control group and receives a plasmid containing the gene encoding the transcription factor and a plasmid containing the test gene without the binding site for the transcription factor.

The production of test gene transcripts or the amount of activity of the relevant protein is measured; if mRNA synthesis from the test gene or the amount of activity of the relevant protein is greater than that of the control gene, the transcription factor is a positive regulator of transcription. If test gene mRNA synthesis or the amount

of activity of the relevant protein is less than that of the control, the transcription factor is a negative regulator of transcription.

Optionally, the assay may include a transfection efficiency control plasmid. This plasmid expresses a gene product independent of the test gene, and the amount of this gene product indicates roughly how many cells are taking up the plasmids and how efficiently the DNA is being introduced into the cells.

A novel chimeric protein has been designed and constructed, as further described in the Examples. This protein, a zinc finger-homeodomain chimeric protein (ZFHD1), has been shown to have DNA-binding specificity in that it binds a composite DNA sequence which includes the sequences individually bound by the components of the chimera; the chimeric protein also weakly binds one of the sequences bound by the individual components of the chimeric protein (the Oct-1 site). As further described in the Examples, a transcription factor (ZFHD1-VP16) was also constructed and shown to have activity in vivo.

To design an appropriate molecule, computer modeling studies were utilized to determine the orientation and linkage of potentially useful DNA-binding domains (see Example 1). Computer modeling studies allowed manipulation and superimposing of the crystal structures of Zif268 and Oct-1 protein-DNA complexes. This study yielded two arrangements of the domains which appeared to be suitable for use in a chimeric protein. In one alignment, the carboxyl-terminal region of zinc finger 2 was 8.8 Å away from the amino-terminal region of the homeodomain, suggesting that a short polypeptide could connect these domains. In this model, the chimeric protein would bind a hybrid DNA site with the sequence 5'-AAATNNTGGGCG-3' (SEQ ID NO: 18). The Oct-1 homeodomain would recognize the AAAT subsite, zinc finger 2 would recognize the TGG

subsite, and zinc finger 1 would recognize the GCG subsite. No risk of steric interference between the domains was apparent in this model. This arrangement was used in the work described below and in the Examples.

The second plausible arrangement would also have a short polypeptide linker spanning the distance from zinc finger 2 to the homeodomain (less than 10 Å); however, the subsites are arranged so that the predicted binding sequence is 5'-CGCCCANNAAAT-3' (SEQ ID NO: 19). This arrangement was not explicitly used in the work described below, although the flexibility of the linker region may also allow ZFHD1 to recognize this site.

After selecting a suitable arrangement, construction of the corresponding molecule was carried out. Generally, sequences may be added to the chimeric protein to facilitate expression, detection, purification or assays of the product by standard methods. A glutathione S-transferase domain (GST) was attached to ZFHD1 for these purpose (see Example 2).

The consensus binding sequence of the chimeric protein ZFHD1 was determined by selective binding studies from a random pool of oligonucleotides. The oligonucleotide sequences bound by the chimeric protein were sequenced and compared to determine the consensus binding sequence for the chimeric protein (see Example 3 and Figure 1).

After four rounds of selection, 16 sites were cloned and sequenced (SEQ ID NOS: 1-16, Figure 1B). Comparing these sequences revealed the consensus binding site 5'-TAATTANGGGNG-3' (SEQ ID NO: 17). The 5' half of this consensus, TAATTA, resembled a canonical homeodomain binding site TAATNN (Laughon, (1991)), and matched the site (TAATNA) that is preferred by the Oct-1 homeodomain in the absence of the POU-specific domain (Verrijzer et al., EMBO J. 11:4993 (1992)). The 3' half of the consensus, NGGGNG,

resembled adjacent binding sites for fingers 2 (TGG) and 1 (GCG) of Zif268.

Binding studies were performed in order to determine the ability of the chimeric protein ZFHD1 to distinguish the consensus sequence from the sequences recognized by the components of the chimera. ZFHD1, the Oct-1 POU domain (containing a homeodomain and a POU-specific domain), and the three zinc fingers of Zif268 were compared for their abilities to distinguish among the Oct-1 site 5'-ATGCAAATGA-3' (SEQ ID NO: 20), the Zif268 site 5'-GCGTGGGCG-3' and the hybrid binding site 5'-TAATGATGGGCG-3' (SEQ ID NO: 21). The chimeric protein ZFHD1 preferred the optimal hybrid site to the octamer site by a factor of 240 and did not bind to the Zif site. The POU domain of Oct-1 bound to the octamer site with a dissociation constant of 1.8×10^{-10} M under the assay conditions used, preferring this site to the hybrid sequences by factors of 10 and 30, and did not bind to the Zif site. The three zinc fingers of Zif268 bound to the Zif site with a dissociation constant of 3.3×10^{-10} M, and did not bind to the other three sites. These experiments show that ZFHD1 binds tightly and specifically to the hybrid site and displayed DNA-binding specificity that was clearly distinct from that of either of the original proteins.

In order to determine whether the novel DNA-binding protein could function in vivo, ZFHD1 was fused to a transcriptional activation domain to generate a transcription factor, and transfection experiments were performed (see Example 5). An expression plasmid encoding ZFHD1 fused to the carboxyl-terminal 81 amino acids of the Herpes Simplex Virus VP16 protein (ZFHD1-VP16) was co-transfected into 293 cells with reporter constructs containing the SV40 promoter and the firefly luciferase gene (Figure 3). To determine whether the chimeric protein could specifically regulate gene expression, reporter

constructs containing two tandem copies of either the ZFHD1 site 5'-TAATGATGGGCG-3' (SEQ ID NO: 21), the octamer site 5'-ATGCAAATGA-3' (SEQ ID NO: 20) or the Zif site 5'-GCGTGGGCG-3' inserted upstream of the SV40 promoter were tested. When the reporter contained two copies of the ZFHD1 site, the ZFHD1-VP16 protein stimulated the activity of the promoter in a dose-dependent manner. Furthermore, the stimulatory activity was specific for the promoter containing the ZFHD1 binding sites. At levels of protein which stimulated this promoter by 44-fold, no stimulation above background was observed for promoters containing the octamer or Zif sites. Thus, ZFHD1 efficiently and specifically recognized its target site in vivo.

Utilizing the above-described procedures and known DNA-binding domains, other novel chimeric transcription factor proteins can be constructed. These chimeric proteins can be studied as disclosed herein to determine the consensus binding sequence of the chimeric protein. The binding specificity, as well as the in vivo activity, of the chimeric protein can also be determined using the procedures illustrated herein. Thus, the methods of this invention can be utilized to create various chimeric proteins from the domains of DNA-binding proteins.

Chimeric DNA-binding proteins of the present invention can be used as transcription factors to turn on or increase the transcription of a gene and thereby increase expression of the gene product. The gene is selected as a result of insufficient expression of its gene product; this insufficient expression may result from a faulty DNA sequence which does not allow binding of the naturally-occurring transcription factor, or from a mutation in or absence of a protein which regulates the gene. This method might be used in any case in which increased transcription of a gene or increased production of a particular RNA species is desired.

For instance, a synthetic DNA sequence recognized by a chimeric transcription factor might be inserted into the promoter of a selected gene, present in a cell as obtained, in order to turn on or cause expression of the gene. DNA encoding the appropriate chimeric transcription factor, comprising at least two binding domains and at least one activation domain, can be incorporated into a construct which is introduced into the cell, where it is expressed. The chimeric protein recognizes and binds to its composite DNA sequence in the promoter, thereby activating transcription of the gene and expression of the gene product.

Alternatively, a gene may not be expressed in a cell or may be expressed at less than normal levels, as a result of a reduced level of, or absence of, an appropriate naturally-occurring transcription factor for recognition of the target DNA sequence in the promoter of a gene. A chimeric transcription factor can be designed to recognize the target sequence by combining DNA-binding domains with at least one activation domain. The chimeric protein can be expressed from a construct which may include an inducible promoter (e.g., the metallothionein promoter). When expressed in the cell, the chimeric protein recognizes and binds the native DNA sequence, thereby activating transcription of the gene.

Novel DNA-binding proteins of the present invention can also be used as transcriptional repressors. For instance, a chimeric transcriptional repressor can be used to control the expression of disease-causing gene products or decrease expression of a product which is overproduced or overactive. In one embodiment, a synthetic DNA sequence recognized by a chimeric DNA-binding protein can be inserted in the promoter of a selected gene. The appropriate chimeric protein, comprising at least two DNA-binding domains, can be incorporated into a construct and

expressed in the cell. The chimeric protein will recognize and bind its composite sequence, thereby interfering with transcription of the gene. Alternatively, the chimeric protein may also comprise a repression domain which acts to negatively regulate transcription of the gene upon binding of the repressor.

In another embodiment, the composite sequence recognized by the chimeric transcription repressor may exist in the promoter of a gene in a cell as obtained. The separate portions of this sequence may normally be recognized by different DNA-binding proteins; the chimeric repressor, comprising at least two DNA-binding domains, recognizes the composite sequence and binds to it, thereby inhibiting binding of the naturally-occurring transcription factor and negatively regulating transcription of the gene. The chimeric repressor may optionally comprise a repression domain which acts to negatively regulate transcription.

Alternatively, the DNA sequences in proximity to a selected gene can be inspected for potential binding sites for individual DNA-binding domains, irrespective of whether any transcription factors bind to those sequences in the normal regulation of the selected gene. Sequences that resemble composite binding sites can be used as targets for the design of chimeric transcription factors which could bind to these native DNA sequences.

The chimeric DNA-binding proteins of the present invention can also be used to construct novel restriction endonucleases. A domain with endonuclease activity (a cleavage domain) can be included in the chimeric protein to cleave the DNA adjacent to the DNA bound by the chimeric protein. For example, the C-terminal cleavage domain of Fok I endonuclease, which has nonspecific DNA-cleavage activity (Li et al., Proc. Natl. Acad. Sci. USA 89:4275-4279 (1992)), can be coupled to the DNA-binding domains of the chimeric protein to generate a new restriction

endonuclease which binds a composite DNA sequence and cleaves DNA adjacent to the binding site of the chimeric protein.

Site-specific restriction enzymes can also be linked to other DNA-binding domains to generate endonucleases with very strict sequence requirements. The chimeric DNA-binding proteins can also be fused to other domains that would control the stability, association and subcellular localization of the new proteins.

Chimeric proteins of the present invention also have utility for manipulation of gene expression for protein production, for RNA production, and for the regulation of gene expression in transgenic animals. For instance, the DNA sequence recognized by the chimeric protein can be inserted into the promoter of a gene of interest. A chimeric transcription factor, i.e., a chimeric protein comprising at least two binding domains from different binding proteins and at least one activation domain, can be incorporated into a construct and expressed under the control of an inducible promoter. When the inducible promoter is turned on, the chimeric protein is expressed and binds to its recognized composite sequence in the promoter of the gene, thereby activating transcription of the gene and expression of the gene product. When the inducible promoter is not turned on, the chimeric protein is not expressed and transcription of the gene of interest is not activated.

The invention will be further illustrated by the following non-limiting examples:

EXAMPLES

The following is a description of design and construction of a chimeric DNA-binding protein, identification of a consensus nucleic acid sequence bound by the chimeric protein, assessment of the binding

specificity of the chimeric protein and demonstration of its *in vivo* activity. The teachings of references cited herein are hereby incorporated by reference.

Example 1: Computer Modeling

Computer modeling studies (PROTEUS and MOGLI) were used to visualize how zinc fingers might be fused to the Oct-1 homeodomain. The known crystal structures of the Zif268-DNA (Pavletich and Pabo, *Science* 252:809 (1991)) and Oct-1-DNA (Klemm, *et al.*, *Cell* 77:21 (1994)) complexes were aligned by superimposing phosphates of the double helices in several different orientations. This study yielded two arrangements which appeared to be suitable for use in a chimeric protein.

Each model was constructed by juxtaposing portions of two different crystallographically determined protein-DNA complexes. Models were initially prepared by superimposing phosphates of the double helices in various registers and were analyzed to see how the polypeptide chains might be connected. Superimposing sets of phosphates typically gave root mean squared distances of 0.5-1.5 Å between corresponding atoms. These distance gave some perspective on the error limits involved in modeling, and uncertainties about the precise arrangements were one of the reasons for using a flexible linker containing several glycines.

In one alignment, the carboxyl-terminal region of zinc finger 2 was 8.8 Å away from the amino-terminal region of the homeodomain, suggesting that a short polypeptide linker could connect these domains. In this model, the chimeric protein would bind a hybrid DNA site with the sequence 5'-AAATNNTGGGCG-3' (SEQ ID NO: 18). The Oct-1 homeodomain would recognize the AAAT subsite, zinc finger 2 would recognize the TGG subsite, and zinc finger 1 would recognize the GCG subsite. No risk of steric interference between the domains was apparent in this model.

The second plausible arrangement would also have a short polypeptide linker connecting zinc finger 2 to the homeodomain (a distance of less than 10 Å); however, the subsites are arranged so that the predicted binding sequence is 5'-CGCCCANNAAAT-3' (SEQ ID NO: 19). This model was not explicitly used in the subsequent studies, although it is possible that the flexible linker will also allow ZFHD1 to recognize this site.

Example 2: Construction of a Chimeric Protein

The design strategy was tested by construction of a chimeric protein, ZFHD1, that contained fingers 1 and 2 of Zif268, a glycine-glycine-arginine-arginine linker, and the Oct-1 homeodomain (Figure 1A). A fragment encoding Zif268 residues 333-390 (Christy et al., Proc. Natl. Acad. Sci. USA 85:7857 (1988)), two glycines and the Oct-1 residues 378-439 (Sturm et al., Genes & Development 2:1582 (1988)) was generated by polymerase chain reaction, confirmed by dideoxysequencing, and cloned into the BamHI site of pGEX2T (Pharmacia) to generate an in-frame fusion to glutathione S-transferase (GST). The GST-ZFHD1 protein was expressed by standard methods (Ausubel et al., Eds., Current Protocols in Molecular Biology (John Wiley & Sons, New York, 1994), purified on Glutathione Sepharose 4B (Pharmacia) according to the manufacturer's protocol, and stored at -80°C in 50 mM Tris pH 8.0, 100 mM KCl, and 10% glycerol. Protein concentration was estimated by densitometric scanning of coomassie-stained SDS PAGE-resolved proteins using bovine serum albumin (Boehringer-Mannheim Biochemicals) as standard. The DNA-binding activity of this chimeric protein was determined by selecting binding sites from a random pool of oligonucleotides.

Example 3: Consensus Binding Sequences

The probe used for random binding site selection contained the sequence 5'-GGCTGAGTCTGAACGGATCCN_{25}CCTCGAG ACTGAGCGTCG-3' (SEQ ID NO: 22). Four rounds of selection were performed as described in Pomerantz and Sharp, Biochemistry 33:10851 (1994), except that 100 ng poly[d(I-C)]/poly[d(I-C)] and 0.025% Nonidet P-40 were included in the binding reaction. Selections used 5 ng randomized DNA in the first round and approximately 1 ng in subsequent rounds. Binding reactions contained 6.4 ng of GST-ZFHD1 in round 1, 1.6 ng in round 2, 0.4 ng in round 3 and 0.1 ng in round 4.

After four rounds of selection, 16 sites were cloned and sequenced (SEQ ID NOS: 1-16, Figure 1B). Comparing these sequences revealed the consensus binding site 5'-TAATTANGGGNG-3' (SEQ ID NO: 17). The 5' half of this consensus, TAATTA, resembled a canonical homeodomain binding site TAATNN (Laughon, (1991)), and matched the site (TAATNA) that is preferred by the Oct-1 homeodomain in the absence of the POU-specific domain (Verrijzer et al., EMBO J. 11:4993 (1992)). The 3' half of the consensus, NGGGNG, resembled adjacent binding sites for fingers 2 (TGG) and 1 (GCG) of Zif268. The guanines were more tightly conserved than the other positions in these zinc finger subsites, and the crystal structure shows that these are the positions of the critical side chain-base interactions (Pavletich and Pabo (1991)).

The consensus sequence of ZFHD1 was determined (5'-TAATTANGGGNG-3', SEQ ID NO: 17), but because of the internal symmetry of the TAATTA subsite this sequence was consistent with the homeodomain binding in either of two orientations (Figure 1C, compare mode 1 and mode 2). The second arrangement (Figure 1C, mode 2), in which the critical TAAT is on the other strand and directly juxtaposed with the zinc finger (TGGGCG) subsites, was considered unlikely since modeling suggested that this

arrangement required a linker to span a large distance between the carboxyl-terminal region of finger 2 and the amino-terminal region of the homeodomain.

To determine how the homeodomain binds to the TAATTA sequence in the 5' half of the consensus, ZFHD1 was tested for binding to probes (5'-TAATGATGGGCG-3', SEQ ID NO: 21, and 5'-TCATTATGGGCG-3', SEQ ID NO: 23) designed to distinguish between these orientations. ZFHD1 bound to the 5'-TAATGATGGGCG-3' probe with a dissociation constant of 8.4×10^{-10} M, and preferred this probe to the 5'-TCATTATGGGCG-3' probe by a factor of 33. This suggests that the first four bases of the consensus sequence form the critical TAAT subsite that is recognized by the homeodomain and that ZFHD1 binds as predicted in the model shown in mode 1 of Figure 1C.

Example 4: Novel Specificity

ZFHD1, the Oct-1 POU domain (containing a homeodomain and a POU-specific domain, Pomerantz et al., Genes & Development 6:2047 (1992)) and the three zinc fingers of Zif268 (obtained from M. Elrod-Erickson) were compared for their abilities to distinguish among the Oct-1 site 5'-ATGCAAATGA-3' (SEQ ID NO: 20), the Zif268 site 5'-GCGTGGGCG-3' and the hybrid binding site 5'-TAATGATGGGCG-3' (SEQ ID NO: 21). DNA-binding reaction contained 10 mM Hepes (pH 7.9), 0.5 mM EDTA, 50 mM KCl, 0.75 mM DTT, 4% Ficoll-400, 300 μg/ml of bovine serum albumin, with the appropriate protein and binding site in a total volume of 10 μl. The concentration of binding site was always lower than the apparent dissociation constant by at least a factor of 10. Reactions were incubated at 30°C for 30 minutes and resolved in 4% nondenaturing polyacrylamide gels. Apparent dissociation constants were determined as described in Pomerantz and Sharp, Biochemistry 33:10851 (1994). Probes were derived by cloning the following

fragments into the Kpn I and Xho I sites of pBSKII+ (Stratagene) and excising the fragment with Asp718 and Hind III: 5'-CCTCGAGGTCATTATGGGCGCTAGGTACC-3' (SEQ ID NO: 24), 5'-CCTCGAGGCGCCCATCATTACTAGGTACC-3' (SEQ ID NO: 25), 5'-CCTCGAGGCGCCCACGCCTAGGTACC-3' (SEQ ID NO: 26), 5'-CCTCGAGGTCATTTGCATACTAGGTACC-3' (SEQ ID NO: 27).

The GST-ZFHD1 protein was titrated into DNA-binding reactions containing the probes listed at the top of each set of lanes in Figure 2. Lanes 1, 6, 11 and 16 contained the protein at 9.8×10^{-11} M, and protein concentration was increased in 3-fold increments in subsequent lanes of each set. The chimeric protein ZFHD1 preferred the optimal hybrid site to the octamer site by a factor of 240 and did not bind to the Zif site.

The Oct-1-POU protein was titrated into DNA-binding reactions as with ZFHD1, but lanes 1, 6, 11 and 16 contained the protein at 2.1×10^{-12} M. The POU domain of Oct-1 bound to the octamer site with a dissociation constant of 1.8×10^{-10} M, preferring this site to the hybrid sequences by factors of 10 and 30, and did not bind to the Zif site.

A peptide containing Zif fingers 1, 2 and 3 was titrated into DNA-binding reactions as with ZFHD1 and the Oct-1-POU protein with lanes 1, 6, 11 and 16 containing the peptide at 3.3×10^{-11} M. The three fingers of Zif268 bound to the Zif site with a dissociation constant of 3.3×10^{-10} M, and did not bind to the other three sites. These experiments show that ZFHD1 binds tightly and specifically to the hybrid site and displayed DNA-binding specificity that was clearly distinct from that of either of the original proteins.

Example 5: In Vivo Activity

ZFHD1 was fused to a transcriptional activation domain, and transfection experiments were used to determine

whether the novel DNA-binding protein could function in vivo. An expression plasmid encoding ZFHD1 fused to the carboxyl-terminal 81 amino acids of the Herpes Simplex Virus VP16 protein (ZFHD1-VP16) was co-transfected into 293 cells with reporter constructs containing the SV40 promoter and the firefly luciferase gene (Figure 3). The 293 cells were co-transfected with 5 μg of reporter vector, 10 μg of expression vector, and 5 μg of pCMV-hGH used as an internal control. The reporter vectors contained two tandem copies of either the ZFHD1 site (TAATGATGGGCG), the Oct-1 site (ATGCAAATGA), the Zif site (GCGTGGGCG) or no insert.

The ZFHD1-VP16 expression vector was constructed by cloning a fragment encoding ten amino acid polypeptide epitope MYPYDVPDYA; ZFHD1; and VP16 residues 399-479 (Pellett et al., Proc. Natl. Acad. Sci. USA 82:5870 (1985)) into the Not I and Apa I sites of Rc/CMV (Invitrogen). Reporter vectors were constructed by cloning into the Xho I and Kpn I sites of pGL2-Promoter (Promega) the following fragments: 5'-GGTACCAGTATGCAAATGACTGCAGTATGCAAATGACCTCGAG-3' (SEQ ID NO: 28), 5'-GGTACCAGGCGTGGGCGCTGCAGGCGTGGGCGCCTCGAG-3' (SEQ ID NO: 29), 5'-GGTACCAGTAATGATGGGCGCTGCAGTAATGATGGGCGCCTCGAG-3' (SEQ ID NO: 30).

The 293 cells were transfected using calcium phosphate precipitation with a glycerol shock as described in Ausubel et al., Eds., Current Protocols in Molecular Biology (John Wiley & Sons, New York, (1994). Quantitation of hGH production was performed using the Tandem-R HGH Immunoradiometric Assay (Hybritech Inc., San Diego, CA) according to the manufacturer's instructions. Cell extracts were made 48 hours after transfection and luciferase activity was determined using 10 μl of 100 μl total extract/10 cm plate and 100 μl of Luciferase Assay Reagent (Promega) in a ML2250 Luminometer (Dynatech Laboratories, Chantilly, VA) using the enhanced flash program and integrating for 20 seconds with no delay. The

level of luciferase activity obtained, normalized to hGH production, was set to 1.0 for the co-transfection of Rc/CMV with the no-insert reporter pGL2-Promoter.

To determine whether the chimeric protein could specifically regulate gene expression, reporter constructs containing two tandem copies of either the ZFHD1 site 5'-TAATGATGGGCG-3', the octamer site 5'-ATGCAAATGA-3', or the Zif site 5'-GCGTGGGCG-3' inserted upstream of the SV40 promoter were tested. When the reporter contained two copies of the ZFHD1 site, the ZFHD1-VP16 protein stimulated the activity of the promoter in a dose-dependent manner. Furthermore, the stimulatory activity was specific for the promoter containing the ZFHD1 binding sites. At levels of protein which stimulated this promoter by 44-fold, no stimulation above background was observed for promoters containing the octamer or Zif sites. Thus, ZFHD1 efficiently and specifically recognized its target site <u>in vivo</u>.

<u>Equivalents</u>

Those skilled in the art will recognize, or be able to ascertain, using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following claims.